DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5365

ISSN No. 0976-5697

Volume 9, No. 1, January-February 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

BIGDATA AND PREDICTIVE ANALYTICS IN THE ELECTION CAMPAIGN

S.Jyothi Mani
M. Tech Student, Department of CSE,
SSJ Engineering College
Hyderabad, Telangana, India

A. Ravi Kumar
Associate Professor, Department of CSE
SSJ Engineering College
Hyderabad, Telangana, India

Abstract: This electronic document is a "live" template. The various components of your paper [title, text, heads, etc.] are already defined on the style sheet, as illustrated by the portions given in this document. Do not use special characters, symbols, or math in your title or abstract. The authors must follow the instructions given in the document for the papers to be published. You can use this document as both an instruction set and as a template into which you can type your own text.

Keywords: Presidential election, predictive analysis, social media

I. INTRODUCTION

Every four years, United States (US) held political event to elect a new president. This event known as US presidential election. The process to run the US presidential election is called electoral college. This event eagerly anticipated not only by US citizens, but also people around the world [1]. The popularity of predicting the US presidential election has been growing, especially in the academic realms [2]. Internet provides enormous data about any kind of topic. Previous research found that gathering data from internet can prove extremely useful for certain domains, including politics [3]. Social media is a part of internet, and it contributes most data in internet. Social media generates large-scale data that shown in a form of millions of users [4]. In this paper, we suggested to use social media as dataset, because social media is widely accessible and up to date [5]. Regarding the use of social media as dataset to predicting the winner of US Presidential election, no one can predict the real intention of user made the post about criticize, praise, or neutral about presidential candidates [6]. This paper proposes a method to predict the winning party or candidate in US presidential election in November 8th, 2016. The data gathered from social media will be processed in four phases: pre-processing, sentiment analysis to classify the sentiment of tweets by using Binary Multinomial Naïve Bayes Classifier, sentiments aggregation to collect the votes, and implementation of Electoral College to predict the chosen party or candidate. Data are collected from Twitter REST API by applying queries about parties and candidates. It must be written in English. However, there is a problem of using social media as dataset that can make data analysis more complicated [7]. Sometimes user wrote their post in daily structured language (e.g. emoticons, slangs, and abbreviation), since the proposed method analyze the data using textual analysis, it can lead to ambiguous extracted information [8]. As a limitation, the proposed method will only get the meaning of abbreviations only. Also, the proposed method only uses 'winner take all' basis for calculating the electors for all states (all electoral votes in a state will be given to candidate which get the majority).

II. LITERATURE REVIEW

Previously, there are some researches have been conducted to predict the US Presidential election, e.g. [1], [2], [4]. Some are using sentiment analysis to represents the casted vote. Each of them were using different data source, methods, and models. Measuring the popularity of US Presidential candidates [1] might be a way to predict the US Presidential election, because it can represent the users interest about the candidates. The candidates analyzed in this researches are Barack Obama and Mitt Romney from US Presidential election 2012. This paper use Web 2.0 (i.e. Blogger, Google News, Twitter, Myspace, YouTube, and Facebook) as a data store, because it contains large amount of information from different users. One way to collect the linguistic dataset is by crawling the Web 2.0 contents. There are three steps of method used in this research: pre-processing, sentiment analysis using SentiWordNet, aggregate by candidate, and visualization presents the popularity graphs of the candidates. Online search traffic can be a data source to predict US Presidential elections [2]. Online search is an information that presumes the searcher knowledge and motivation. Data are collected from Google Trends by using presidential candidate queries and issue queries, because it's freely available for download.

This paper use baseline model to predict the percentage popular vote of each party in each state. The problem is the search queries does not provide information about political ideology, age, gender, and user behavior. Furthermore, a query can show many unrelated topics that do not fit with presidential election. Research [4] provides a systematic link between data grabbed from social media with real-world political behavior. Dataset are collected from Twitter, the Federal Election Commission, and the US Census Bureau. After dataset are collected, this research conducting some variables to find vote share for each district, then analyzing the variables by estimate the effect of Twitter on electoral outcomes using three ordinary least squares algorithm (OLS) models. This paper proves that social media can be a reliable data source about political behavior.

III. THEORETICAL BASIS

A. Sentiment Analysis: Sentiment analysis is a technique for analyzing a large number of documents to obtain writer's sentiment on a topic [10]. Sentiment analysis uses Natural Language Processing (NLP) to collect opinion and examine opinion or sentiment words [9]. Based on [10], there are 2 important tasks in sentiment analysis. First, identify the opinion targets (aspects, entity, and topics). Second, construct the opinion lexicon (e.g. good, excellent, etc.). There are several methods to classify the sentiment of text, e.g. lexicon-based methods and learning based methods. The example of lexicon-based methods is SentiWordNet and AFINN-111. Both SentiWordNet and AFINN-111 are a text lexicon that contains keyword or synset with its positive and negative sentiment score. Learning-based methods for sentiment analysis is a method that need to train its algorithm and use the knowledge to classify the sentiment.

The example of learning-based methods is Vector Space Model (VSM) and Naïve Bayes Classifier. There is various model of Naïve Bayes Classifier, e.g. Multinomial Naïve Bayes, Binary Multinomial Naïve Bayes, and Bernoulli Naïve Bayes.

B. Electoral College: Electoral College is a process to decide a president by gathering 538 (435 representatives, 100 senators, and 3 electors given to the District of Columbia) votes from electors for each state. The electors are selected by political parties and assigned into 50 states equal to its congressional representation [3]. Each state has different numbers of electors. Electors can vote their party or candidate if those electors win the popularity vote in its state. The electoral votes will be summed up to decide the winning party and candidate. Though a candidate lose the popular votes in a state, that candidate still have a chance to win the electoral college (happened to George W. Bush in 2000) [4], [5]. The proposed method uses 'winner take all' basis to decide the winning party or candidate. `Winner take all` will give all the casted electoral votes to the party or candidate which get the majority votes.

IV. RESEARCH METHODOLOGY

This paper proposed two stages in research methodology, i.e. data collection and implementation. The description of both stages in research methodology are explained bellow. A. Data Collection In this research, data are grabbed from Twitter REST API and must be written in English. The collected data must contain tweet status/text (excluding retweets), post time, username, and user location. Tweet status should mention any keyword about parties or candidates that participating in the US presidential election 2016 (e.g. republican, democrat, Hillary Clinton, Donald Trump). The tweets will represent the vote casted for its keyword related candidate. User location must be in one of 51 states that listed in Electoral College. Corpus is needed to do learning-based sentiment analysis method such as Binary Multinomial Naïve Bayes. This research using Sentiment140 tweet corpus [6] that contains a total of 1,600,000 data train (divided into 800,000 data training for each positive and negative sentiment) and 497 data test (divided into 181 positive data, 177 negative data, and 139 neutral data). To minimize the processing time yet still generating good results, this research limits the data training used to 10,000 positive data and 10,000 negative. As for the data test, 181 positive data and 177 negative data (because data training only provide positive and negative data) were analyzed to test the performance of sentiment analysis method. Abbreviations dictionary is needed to expand the abbreviations found in tweet status. The dictionary must contain the acronym and its meaning, e.g. LOL means laughing out loud, B/C means because, GF means girlfriend, CMIIW means correct me if I'm wrong. This dictionary are gathered by from [7] content. Acronyms that only have a character (e.g. u for you, d for the, c for see) and have the same characters with English word (e.g. HOPE for Have Only Positive Expectation) do not included. Moreover, the acronyms that have multiple meaning (e.g. LML can means Laughing Mad Loud and Love My Life) do not included to prevent ambiguity. Contraction dictionary is an array that contains list of contractions (e.g. I'm means I am, don't means do not, won't means will not, etc.).

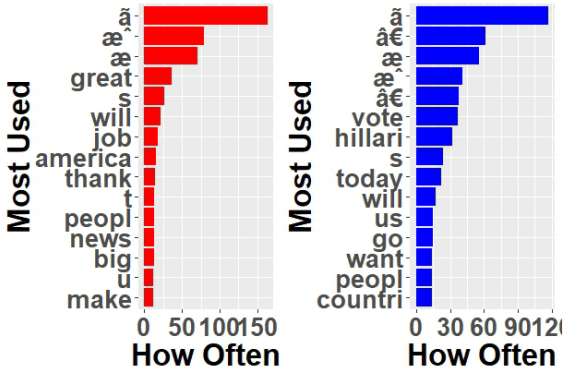

Fig 1: Polls for parties.

The contraction dictionary are made by referring from Cambridge Dictionary [8]. Stop words are common words that carry less important meaning than keywords [9]. This research use the free English stop words list downloaded from [2]. B. Implementation The implementation consists of pre-processing, sentiment analysis, aggregation, and implementing electoral college to predict the winning candidate or party. Fig. 1 shows the part of implementation in sequence.

Pre-processing: Tweets status must be pre-processed to get the valuable keywords or tokens. The HTML elements, URLs, and mentions (e.g. @username) should be removed from tweets status. Then, the 'clean' tweets status are tokenized into array of tokens or keywords (bag of words). All contractions and abbreviations found in the tokens will be expanded in accordance with its meaning as it exists in the abbreviations and contractions dictionary. Afterward, remove the non-alphanumerical characters in tokens and remove every tokens that contained in stop words list to return the most relevant result [9]. The remaining tokens will be reduced to its base word (stem) using Porter Stemmer algorithm. This stage will end after the system generates the unigram and bigram from the tokens. This research only limits the n-gram to unigram and bigram to maximize the analysis process results, yet the process time won't be too long.

Aggregation: The purpose of this stage is to aggregate the sentiments of tweets to decide the winning electors for each state. The tweets sentiment will be used to represent the vote casted by the user who wrote the tweet status. If the sentiment of a tweet is positive, then the vote is given to the mentioned party or candidate. But, if the sentiment of a tweet is negative,

the vote is given to the opposite of mentioned party or candidate. The electors chosen in a state are decided by popularity vote. The winning electors are the one who has the most votes.

Implementing Electoral College: The electors who win the popularity vote in its state, must cast their electoral votes to the party that appointed them. The number of electoral votes are distributed according to the allocation of electors for each state. The party or candidate who get the most electoral votes is predicted to be the winner of US presidential election 2016.

V. RESULTS AND DISCUSSION

This section explains the test results conducted by author by using the proposed method mentioned in research methodology. Researchers have made an application to test pre-processing and sentiment analysis.

As illustrated in Fig. 1, the prediction result shows that Donald Trump (Republican) win the election with more user tweets ratings, since Hillary Clinton (Democrat) only got 219 electoral votes. It happened because the data real collected from Twitter. The tweets gets updated each time comes up with different results.

VI. CONCLUSION

This paper proposes a method for predicting the US presidential election by using two stages: data collection and implementation. Previously, several research has been conducted to predict the US presidential election. The proposed method is created by referring the previous researches and adding some value (i.e. data collection technique, abbreviations and contractions dictionary, and the implementation of electoral college) to make the prediction more accurate and also adjust to the actual situation. There are some deficiencies found during research, i.e. there is a possibility that the tweets used as dataset were written by same user, in actual situation a person can only cast one vote. The proposed method can't decide the winning party or candidate if the votes is tie, because there are regulations about tiebreaking vote that decided by the senates that cannot be implemented in system. The implementation of electoral college unable to process real time data due to the random value of user location provided from Twitter REST API. For future work, it will be implemented using real time data. The proposed method also can be used as a reference to do similar research about election in other country.

VII. REFERENCES

[1] I. Malinský, Radek; Jelínek, "Sentiment Analysis: Popularity of Candidates for the President of the United States," Proc. World Acad. Sci. Eng. Technol., vol. 72, pp. 1382–1384, 2012.

[2] L. Granka, "Using Online Search Traffic to Predict US Presidential Elections," PS Polit. Sci. Polit., vol. 46, no. 02, pp. 271–279, Apr. 2013.

[3] E. Șt Chifu, T. Șt Leția, B. Budișan, and V. R. Chifu, "Web Harvesting And Sentiment Analysis Of Consumer Feedback," ACTA Tech. NAPOCENSIS Electron. Telecommun., vol. 56, no. 3, 2015.

[4] J. DiGrazia, K. McKelvey, J. Bollen, and F. Rojas, "More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior," PLoS One, vol. 8, no. 11, p. e79449, Nov. 2013.

[5] L. S. L. Lai and W. M. To, "Content Analysis of Social Media : a Grounded Theory Approach," J. Electron. Commer. Res., vol. 16, no. 1, pp. 138–152, 2014.

[6] H. Schoen, D. Gayo-Avello, P. Takis Metaxas, E. Mustafaraj, M. Strohmaier, and P. Gloor, "The power of prediction with social media," Internet Res., vol. 23, no. 5, pp. 528–543, Oct. 2013.

[7] B. Batrinca and P. C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms," AI Soc, vol. 30, pp. 89–116, 2014.

[8] R. Irfan, C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Wang, D. Chen, A. Rayes, N. Tziritas, C.-Z. Xu, A. Y. Zomaya, A. S. Alzahrani, and H. Li, "A survey on text mining in social networks," Knowl. Eng. Rev., vol. 30, no. 02, pp. 157–170, Mar. 2015.

[9] M. Aarti and A. Patil, "Sentiment Analysis for Product Reviews," Int. J. Adv. Res. Comput. Sci., vol. 5, no. 5, pp. 202–204, 2014.

[10] S. Patni and W. M. E. Avinash, "Review Paper on Sentiment Analysis Using Web 2.0 by Classification Method," Int. J. Adv. Res. Comput.